Top2Vec: Distributed Representations of Topics

Year: 2020 | Citations: 424 | Authors: D. Angelov

Abstract

Topic modeling is used for discovering latent semantic structure, usually referred to as topics, in a large collection of documents. The most widely used methods are Latent Dirichlet Allocation and Probabilistic Latent Semantic Analysis. Despite their popularity they have several weaknesses. In order to achieve optimal results they often require the number of topics to be known, custom stop-word lists, stemming, and lemmatization. Additionally these methods rely on bag-of-words representation of documents which ignore the ordering and semantics of words. Distributed representations of documents and words have gained popularity due to their ability to capture semantics of words and documents. We present $\texttt{top2vec}$, which leverages joint document and word semantic embedding to find $\textit{topic vectors}$. This model does not require stop-word lists, stemming or lemmatization, and it automatically finds the number of topics. The resulting topic vectors are jointly embedded with the document and word vectors with distance between them representing semantic similarity. Our experiments demonstrate that $\texttt{top2vec}$ finds topics which are significantly more informative and representative of the corpus trained on than probabilistic generative models.